Weather data errors analysis in solar power stations generation forecasting

Stanislav Eroshenko[1] and Alexandra Khalyasmaa[1*]

[1]Ural Federal University named after the first President of Russia B.N. Yeltsin, Ekaterinburg, 620002 Mira str. 19, Russia

Abstract. The paper presents a short-term forecasting model for solar power stations (SPS) generation developed by the authors. This model is based on weather data and built into the existing software product as a separate short-term forecasting module for the SPS generation. The main problems associated with forecasting the SPS generation on cloudy days were revealed in the framework of authors' research, which is due not to the error of the developed model but to the use of the same learning sample for both solar and cloudy days. This paper contains analysis of the main problems related to the learning sampling, samples pattern, quality and representativeness for forecasting the SPS generation on cloudy days. Besides, the paper includes a calculation example performed for the existing SPS and a detailed analysis of the forecast generation on cloudy days based on the actual weather provider data.

1 Introduction

Today, the main problem for the short-term (day ahead) forecasting of solar power stations (SPS) generation is the necessity to use an accurate and detailed weather forecast, which in turn is not cheap for the customer (SPS owner). In order to save money, the customer often uses the available free weather data, which, according to authors' research, is ultimately insufficient for accurate identifying the weather conditions necessary to solve the presented problem.

It should be noted that the weather data composition is of statistical significance from the point of view of the influence on the transparency factor calculation and allows to establish a reliable statistical relationship between the properties (cloud cover, the altitude of the solar disk) and the response (the solar radiation flux density on a horizontal surface) [1]. The impossibility of identifying the weather conditions leads to a high variance of the response y (transparency factor) with respect to the features under consideration xi (the estimate of cloud cover, the sine of the solar disk altitude angle).

An estimated evaluation shows that, on average, every fifth change in the actual value of the transparency factor does not correspond to a change in the actual cloud cover. At the same time, if in addition to take the forecast cloud cover into account, which is a priori less accurate than the actual one, the error of the initial data increases which introduces a significant error in the calculation results. To increase the accuracy of forecasting on days with precipitation, it is necessary to accumulate appropriate retrospective data and get a separate sample for days with precipitation. It is necessary to clearly distinguish the type of precipitation: drizzle, rain, rain shower, hail, snow, sleet, etc.

2 Short-term SPS forecasting model

The presented methodology is based on calculation of the solar radiation flux density (external illumination) on a horizontal surface for a six-day interval according to the weather provider data for half-hour intervals that is necessary for the regression model development. The methodology is described in detail in [2]. The main stages of the developed methodology implementation are presented below.

2.1 Forecast calculation of the solar radiation flux density incident on a horizontal surface`

The average value of the solar radiation flux density near the ground, incident on the horizontal surface, $\overline{G}$ for each hour of the forecast day is determined by the expression:

$$\overline{G} = \overline{k}_T \cdot \overline{G}_0, \quad (1)$$

where $\overline{G}$ is the average value of the solar power flux density near the ground for each hour of the forecast day, [W/m^2]; $\overline{G}$ is the transparency factor for each hour of the forecast day, [p.u.]; $\overline{G} = \overline{k}_T \cdot \overline{G}_0$ is the average value of the solar radiation flux density at the boundary of the atmosphere for each hour of the forecast day, [W/m^2].

To define a parameter $\overline{G}$ a physical model is used that allows to determine solar power flux density at the boundary of the atmosphere, while only the location of the power station, the number of the day in question and the hour of the day in question are needed. The algorithm

* Corresponding author: stas_ersh@mail.ru

for calculating solar power flux density at the boundary of the atmosphere is described by the following formulas.

- The average value of the solar radiation flux density at the boundary of the atmosphere $\overline{G}_0$ [3]:

$$\overline{G}_0 = \frac{12}{\pi} G_{on} (\cos\varphi \cdot \cos\delta \cdot (\sin\omega_2 - \sin\omega_1) + \frac{\pi(\omega_2 - \omega_1)}{180^0} \cdot \sin\varphi \cdot \sin\delta) \quad (2)$$

To define a parameter k_T a statistical model is used that allows determining the transparency factor for the next day. Retrospective cloud data from previous days, retrospective data on the solar radiation flux density from previous days, forecast cloud data for the next day, data on location of the power station, the number of the day in question, and the hour of the day are needed.

The regression model is described by the following expression:

$$\overline{k_T} = a_1 + b_1 \cdot (\overline{cc})^2 \cdot \sin\overline{\alpha} + c_1 \cdot (\overline{cc})^2 + d_1 \cdot \sin\overline{\alpha}, \quad (3)$$

where $\overline{k_T}$ is the transparency factor, [p.u.]; $\overline{cc}$ is the cloud cover, [p.u.]; $\overline{\alpha}$ is the average value of the solar altitude angle within the defined time step, [degrees]; a_1, b_1, c_1, d_1 are coefficients determined empirically by regression analysis equations, [dimensionless values].

In order to minimize the manual labor of the operational staff of the SPS in all subsequent calculations, it was decided not to implement the estimate of cloud cover cc in its original form, since this requires a complex analysis of retrospective data of considerable depth (up to two years) for the cloud cover TCC and density of the solar radiation flux incident on the horizontal surface of the earth $\overline{G}$ in order to determine the effect of the cloud character on the magnitude $\overline{G}$.

In further calculations, the value TCC is equal to the total cloud cover TCC in relative units [4]:

$$cc = \frac{TCC}{100}, \quad (4)$$

where cc is the cloud cover, [p.u.]; TCC is the total cloud cover, [%].

The transparency factor $\overline{k_T}$ for the previous days is determined by the known values of the solar radiation flux density for the previous days $\overline{G}$, obtained from the data of the local weather station, and the calculated values $\overline{G}_0$ according to the formula:

$$\overline{G} = \overline{k}_T \cdot \overline{G}_0. \quad (5)$$

Regression analysis is performed for certain intervals of angles $\overline{\alpha}$. For each interval, the coefficients of the regression model a_1, b_1, c_1, d_1 are calculated using expression [5-7]:

$$B^T = \left(X^T X\right)^{-1} \left(X^T Y\right), \quad (6)$$

where B^T is the column vector of regression coefficients a_1, b_1, c_1, d_1; X is the matrix of independent variables composed of the corresponding values $(\overline{cc})^2 \cdot \sin\overline{\alpha}$, $(\overline{cc})^2$ and $\sin\overline{\alpha}$; Y is the matrix of dependent variables, composed of the corresponding values $\overline{k_T}$.

2.2 Accuracy analysis of the SPS generation forecasting

Accuracy analysis of the forecast is performed by means of determining the forecast error. In this study, WAPE error on hours and WMAPE, % are evaluated [8,9]:

- Weighted absolute percentage error (APE):

$$WAPE = \frac{\left|\overline{G}^m(t) - \overline{G}^f(t)\right|}{\overline{G}_0(t)} \cdot 100\%, \quad (7)$$

where APE is the absolute percentage error, [%]; $\overline{G}^m(t)$ is the average measured value of the solar power flux density near the ground within a defined time step on the time interval t, [W/m^2]; $\overline{G}^f(t)$ is the average forecast value of the solar power flux density near the ground within a defined time step on the time interval t, [W/m^2]; $\overline{G}_0(t)$ is the value of the solar radiation flux density at the boundary of the atmosphere for a defined time step t, [W/m^2].

- Weighted mean absolute percentage error (WMAPE):

$$WMAPE = \frac{1}{N} \sum_{t=1}^{N} \frac{\left|\overline{G}^m(t) - \overline{G}^f(t)\right|}{\overline{G}_0(t)} \cdot 100\%. \quad (8)$$

Estimating the error with regard to the solar radiation flux density at the boundary of the atmosphere will allow estimating the error relating to the conditionally constant "base" and to show how many percent the error is from the maximum possible solar radiation flux density for a defined time interval [10].

3 Calculation example of short-term forecasting of SPS generation

Calculation example presents calculations of short-term forecasts within the period of 04.10.2017 – 21.10.2017 for SPS-1 using different learning samples. In accordance with the proposed clustering methodology, to improve the accuracy of short-term forecasting for the days under consideration, a new sample is formed for each day reflecting the weather conditions.

As part of the data provided by the weather provider, there are: a cloud cover estimate, %; air temperature, ^{o}C; relative humidity, %; wind speed, m/s. Cloud cover is a key parameter that affects the transparency factor. Thus, it is most expedient to cluster the initial array of retrospective data in accordance with the gradations of cloud cover. The following rules for the formation of blocks of retrospective data were used in the calculations:

- The retrospective includes the values of the corresponding characteristics of x_i and the y response for cloudy days having an average cloud cover within the light day that differs by

no more than 30% from the average cloud cover within the light day in the current forecast day.

- The retrospective includes the values of the corresponding characteristics of x_i and the y response for cloudy days having an average cloud cover within the light day differing by not more than 50% of the average cloud cover within the light day in the current forecast day.
- The retrospective includes the values of the corresponding characteristics of x_i and the y response in the range of the minimum and the maximum actual cloud cover for the light day in the current forecast day.
- The retrospective includes the values of the corresponding characteristics of x_i and the y response for all available range.

An example of calculation results is presented for the two most problematic (cloudy) days on October 17-18, 2017 in Table 1.

3.1. Retrospective analysis

In this research, a 6-day retrospective was used to forecast the SPS generation initially for its program implementation. While using a 6-day retrospective in the framework of the presented algorithm, the obtained error is random since the learning sample to calculate the regression coefficients "slides" in 1-day step. In this case, it may occur that the forecast is performed on a cloudy day based on clear days and alternately, which negatively affects the accuracy of short-term forecasting and leads to unreliable results.

In the framework of previous research, the recommendation to use the 6-day retrospective was due to the following factors:

- The amount of initial data on SPS, which was originally provided by the customer for a period of 2 weeks, containing full information about weather conditions and the solar power flux density measurements at SPS.
- The Customer's requirement to minimize the necessary amount of retrospective data to ensure the possibility of rapid implementing the forecasting system at new photovoltaic stations.

3.2. Regression function coefficients

The coefficients of the regression function are chosen in such a way as to minimize the approximation error of the available retrospective data. In the absence of some combination of xi and the y response in the retrospective data, for example, a negative design transparency factor might be obtained when the cloud cover increases to 0.9 – 1.0, which indicates the necessity to expand and/or refine the learning sample.

Therefore, in order to obtain a stable forecast, the learning sample should contain data within the order of 80% for cloudy days, about 20% for clear days. Even when forecasting the solar power flux density for cloudy days, the presence of clear days in retrospective data is necessary to obtain stable coefficients of the regression function.

3.3. Learning sample expanding

The expansion of the learning sample positively affects the accuracy of short-term forecasting. With averaging of the best results obtained within the period of 04.10.2017 – 21.10.2017, the mean absolute percentage error was about 28%. The mean absolute percentage error in the solar power flux density forecast was 49%. The use of the entire available retrospective on average over the period under review provides an error of 38%

For the period under consideration, a number of "problem" days can be identified, when the mean absolute percentage error can reach 100% or more. This problem is due to atypically low transparency factors for the presented cloud conditions. For example, these are 18.10.2017, 17.10.2017.

The retrospective database provided by the Customer does not contain information that uniquely identifies weather conditions; therefore, additional open sources on the Internet have been analyzed to identify cause-effect relationships. At the same time, the database of actual weather conditions of weather providers A and B was considered.

It is remarkable that open databases of actual weather information are not available for specific geographic coordinates and are associated with the locations of weather stations. The closest to the SPS-1 location meteorological station is the Airport-1 weather station, located at a distance of 45 km from SPS-1.

The meteorological station remoteness from the SPS-1 location does not allow using the weather data archive for forecasting, but it provides an opportunity to assess the meteorological situation. Both weather providers A and B provide the same information on the Airport-1 weather station, but B presents more detailed set of meteorological parameters.

For the forecast day of 18.10.2017, broken clouds with a low cloud height of 500 m are characteristic. For the forecast day of 17.10.2017, broken clouds with a low cloud height of 500 m are typical; the type of clouds is cumulonimbus, precipitation is in the form of rain and rain shower (Table 2).

Presented days are characterized by the presence of dense cloud cover and precipitation. As a rule, SPS are built in areas with the greatest number of sunny days per year, therefore the presence of days with precipitation is atypical and the forecast for these days is built with a significant overestimation of the transparency factor, and as a consequence, of the solar power flux density.

Increasing the accuracy of the forecast was achieved by constructing a separate sample of days with precipitation according to the data presented in the archive of Airport-1 on the portal of weather provider B. Thus, the forecast error was reduced: - for 17.10.2017: from 135% to 100%; - for 18.10.2017: from 196% to 75%.

To increase the accuracy of forecasting on days with precipitation, it is necessary to accumulate an appropriate retrospective of the data and to construct a separate sample for days with precipitation. It is necessary to clearly distinguish the type of precipitation: drizzle, rain, rain shower, hail, snow, sleet, etc.

3.4. Revealing the main errors in the learning sample

The greatest contribution to MAPE for the considered day is provided by the hours when the actual data on cloud cover does not correspond to the measured value of the transparency factor. For example, an increase in cloud cover in percent leads to an increase in the transparency factor or alternately. Verification of the data archive issued by the customer was carried out under the following conditions:

- The change in the transparency factor and cloud cover in r.u. are of the same sign; for example, a decrease in cloud cover corresponds to a decrease in the transparency factor.
- Change of one value by more than 0.1 r.u. relative to another does not lead to a change in the second value.

For a sample of 0-20 degrees for the solar altitude angle above the horizon, the number of discrepancies was 137 from the total sample of 503 values (27.2%). For a sample of 20-40 degrees for the solar altitude angle above the horizon, the number of discrepancies was 56 from the total sample of 295 values (18.9%). An estimated evaluation shows that, on average, every fifth change in the actual value of the transparency factor does not correspond to a change in the actual cloud cover. Moreover, in case of taking the forecast cloud cover into account, which is a priori less accurate than the actual one, the error of the initial data increases that introduces a significant error in the calculation results.

The impossibility of identifying weather conditions leads to a high variance of the response y (transparency factor) with respect to the features under consideration x_i (the estimate of cloud cover, the sine of the solar disk altitude angle). For example, Figures 1−3 show the transparency factor function for the angles of the solar disk position above the horizon of 0-20 degrees in the point cloud of the retrospective data (the regression function coefficients were obtained for all the available retrospective data on SPS-1 for angles of 0-20 degrees: $a_1 = 0.3273$, $b_1 = -1.1467$, $c_1 = -0.1404$, $d_1 = 1,0551$).

Table 1. Results of the short-term forecast calculations for cloudy days

Date	Characteristic	Retrospective			MAPE, %	WMAPE, %	Retrospective	MAPE, %	WMAPE, %
		Sample size, pcs.		Principle of sampling					
18.10.17	Mean cloud cover (0–20): 0.378 Mean cloud cover (20–40): 0.474	0.0 – 20.0	99	+/- 30% of the actual mean cloud cover per day	94.865	15.518	6 days 0.0 – 20.0: ~96 20.0 – 40.0: ~168	196.095	29.085
		20.0 – 40.0	43						
		0.0 – 20.0	148	+/- 50% of the actual mean cloud cover per day	134.862	19.352			
		20.0 – 40.0	93						
		0.0 – 20.0	194	Based on "Broken clouds" criterion with a cloud height of more than 500 m	75.352	11.138			
		20.0 – 40.0	43						
		0.0 – 20.0	247	in the min/max range of the actual cloud cover within the day	127.012	20.049			
		20.0 – 40.0	203						
		0.0 – 20.0	503	all sample values	135.724	20.777			
		20.0 – 40.0	295						
17.10.17	Mean cloud cover (0–20): 0.940 Mean cloud cover (20–40): 0.977	0.0 – 20.0	265	+/- 30% of the actual mean cloud cover per day	168.7	16.11	6 days 0.0 – 20.0: ~96 20.0 – 40.0: ~168	135.681	13.258
		20.0 – 40.0	138						
		0.0 – 20.0	319	+/- 50% of the actual mean cloud cover per day	155.08	14.724			
		20.0 – 40.0	202						
		0.0 – 20.0	110	Based on "Precipitation", "Fog", "Mist" criteria	101.37	9.314			
		20.0 – 40.0	74						
		0.0 – 20.0	170	in the min/max range of the actual cloud cover within the day	118.309	11.507			
		20.0 – 40.0	75						
		0.0 – 20.0	503	all sample values	138,692	13,352			
		20.0 – 40.0	295						

		20.0 – 40.0	295						

Table 2. The actual weather data of weather provider B

Date and time		Description of weather events	Cloud characteristic
18.10.2017	18.10.17 14:30	—	Broken (60-90%) 870 m
	18.10.17 14:00	—	Broken (60-90%) 1200 m
	18.10.17 13:30	—	Broken (60-90%) 750 m
	18.10.17 13:00	—	Broken (60-90%) 960 m
	18.10.17 12:30	—	Broken (60-90%) 900 m
	18.10.17 12:00	—	Broken (60-90%) 840 m
	18.10.17 11:30	—	Broken (60-90%) 750 m
	18.10.17 11:00	—	Broken (60-90%) 570 m
	18.10.17 10:30	—	Broken (60-90%) 510 m
	18.10.17 10:00	—	Broken (60-90%) 510 m
	18.10.17 09:30	—	Broken (60-90%) 630 m
	18.10.17 09:00	—	Broken (60-90%) 780 m
	18.10.17 08:30	—	Broken (60-90%) 750 m
	18.10.17 08:00	—	Broken (60-90%) 750 m
	18.10.17 07:30	—	Broken (60-90%) 600 m
17.10.2017	17.10.17 14:30	—	Continuous (100%) 900 m
	17.10.17 14:00	—	Continuous (100%) 780 m
	17.10.17 13:30	—	Continuous (100%) 660 m
	17.10.17 13:00	—	Continuous (100%) 630 m
	17.10.17 12:30	—	Continuous (100%) 630 m
	17.10.17 12:00	—	Continuous (100%) 600 m
	17.10.17 11:30	—	Continuous (100%) 540 m
	17.10.17 11:00	Light rain	Continuous (100%) 480 m
	17.10.17 10:30	Light rain	Continuous (100%) 480 m
	17.10.17 10:00	Light rain	Continuous (100%) 480 m
	17.10.17 09:30	Light rain	Continuous (100%) 480 m
	17.10.17 09:00	Light rain, mist	Continuous (100%) 480 m
	17.10.17 08:30	Light showers, rain	Broken (60-90%) 600 m, cumulonimbus clouds
	17.10.17 08:00	Light showers, rain	Broken (60-90%) 600 m, cumulonimbus clouds
	17.10.17 07:30	Light showers,	Broken (60-90%) 600 m,
		rain	cumulonimbus clouds
	17.10.17 07:00	Light showers, rain	Scattered (40-50%) 480 m, broken (60-90%) 690 m, cumulonimbus clouds

Fig. 1 and 2 show the range of actual transparency factors for a single cloud cover value in percent. Blue, red and green are point clouds in the neighborhood (± 2.5°) of the solar altitude angles above the horizon of 15, 10 and 5 degrees, respectively. The blue, red and green lines depict the functions of the transparency factor at fixed angles of the solar height above the horizon of 15, 10 and 5 degrees, respectively. So, for example, for the Sun angular altitude at 12.5 – 17.5 degrees, the cloud cover value equal to 0.5 corresponds to the transparency factors equal to 0.06, 0.16, 0.25, 0.31, 0.44, 0.50, 0.61, 0.77, etc. The spread is about 70%. The regression function averages the existing retrospective similar to any statistical approach. For the case under consideration, this means impossibility to consider unique weather characteristics for a number of extreme conditions (snow, hail, rain shower, fog, etc.), but increase in the stability of the forecasting system to data exclusions.

It should be noted that the very system of converting the initial data, provided by the weather provider, from the linguistic form to the numeric one by the Customer also helps to reduce the informative content of the learning sample.

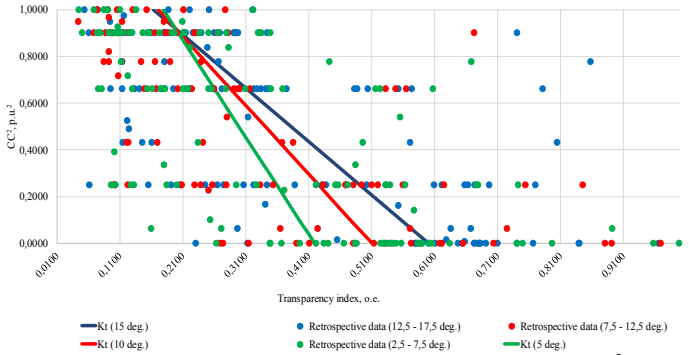

Figure 1. Regression function projection on the axis CC^2

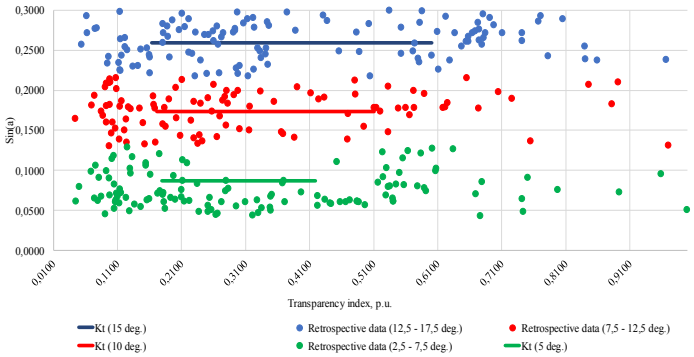

Figure 2. Regression function projection on the axis sinα

4 Conclusion

It is necessary to accumulate additional information on meteorological events and conditions to increase the accuracy of short-term forecasting of SPS generation on cloudy days. It will allow increasing the amount of retrospective and dividing the sampled population into separate samples that will better identify such events as "heavy/light rain", "rain shower", "drizzle", "heavy/light snow", "hail", "fog", "mist", and also classify cloud cover.

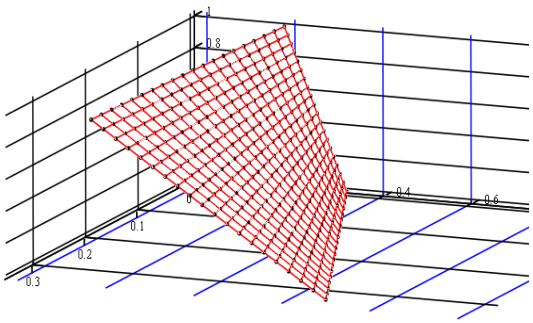

Figure 3. The surface of the transparency factor function for the range of solar altitude angles of 0-20 degrees.

The daily preservation of all possible parameters provided by the weather provider of actual and forecast information (including the original one - the linguistic data provided by the weather provider) is needed for the subsequent improvement of the quality of the SPS generation forecast by reducing the variance of the measured cloud cover values and the calculated transparency factor.

The recommended amount of retrospective allowing accounting the climatic features of the SPS location, as well as clustering of the sampled population into separate samples, is at least 1 year. Provided there are no additional characteristics of weather conditions and events, for short-term forecasting it is recommended:

- To form a retrospective for clear days in the range of an average cloud cover over a day of 0.0 – 0.5 r.u.
- To use the entire available retrospective for cloudy days in the following ratio: 20% of clear days, 80% of cloudy ones.

For new objects (returned to service) it is recommended to use the learning sample according to the rules described above, while ensuring that monthly retrospective record of the data (mostly cloudy days) per sample is made during the current year. The following ratio in the complete sample should be maintained: no more than 20% clear days, not less than 80% of cloudy ones, in addition keeping the total number of pair values (input-output) in the sample of not less than 800 pcs. Record of new retrospective data to the complete sample is made taking into account the recommendations described above, including in its original form - the linguistic data provided by the weather provider.

It is recommended to refine the calculation and display of the forecast error for the short-term SPS generation forecasting and use the APE to estimate the error by hours and the MAPE to estimate the error by day.

References

1. W. Glassley, J. Kleissl. Current state of the art in solar forecasting. California Renewable Energy Collaborative Final Report. 2013.
2. D.A. Snegirev, S.A. Eroshenko, R.T. Valiev, A.I. Khalyasmaa. Algorithmic realization of short-term solar power plant output forecasting. Proceedings of 2017 IEEE 2nd International Conference on Control in Technical Systems, CTS 2017. Pp. 228-231.
3. K.N. Shukla, Saroj Rangnekarb, K. Sudhakar. Comparative study of isotropic and anisotropic sky models to estimate solar radiation incident on tilted surface. Energy Reports 1, pp. 96-103. 2015.
4. W.D. Turner, A. Mujahid. The Estimation of Hourly Global Solar Radiation Using a Cloud Cover Model Developed at Blytheville. American Meteorological Society. 1984, pp. 781-786.
5. M. Paulescu, E. Paulescu, P. Gravila, V. Badescu. Weather Modeling and Forecasting of PV Systems Operation. Springer. 2013.
6. V. Prema, K. Uma Rao. Development of statistical time series models for solar power prediction. Renewable energy. 2015.
7. M. Abuella, B. Chowdhury. Solar Power Probabilistic Forecasting by Using Multiple Linear Regression Analysis. Department of Electrical and Computer Engineering, University of North Carolina, USA. 2015.
8. A. Gondalia, C. Shah. Solar power forecasting analysis of trends in modeling techniques and error minimization mechanism. 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2017, pp. 1860-1866.
9. V. Layanun, S. Suksamosorn, J. Songsiri, "Missing-data imputation for solar irradiance forecasting in Thailand," 2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Kanazawa, 2017, pp. 1234-1239.
10. M. Alanazi, M. Mahoor, A. Khodaei. Two-stage hybrid day-ahead solar forecasting. 2017 North American Power Symposium (NAPS), Morgantown, WV, 2017, pp. 1-6.